

METHODS AND APPARATUS FOR INFORMATION STORAGE AND RETRIEVAL USING A HASHING TECHNIQUE WITH EXTERNAL CHAINING AND ON-THY REMOVAL OF EXPIRED DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

Not Applicable

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

Not Applicable

REFERENCE TO A MICROFICHE APPENDIX

Not Applicable

BACKGROUND OF THE INVENTION

This invention relates to information storage and retrieval systems, and, more particularly, to the use of hashing techniques in such systems.

Information or data stored in a computer-controlled storage mechanism can be retrieved by searching for a particular key value in the stored records. The stored record with a key matching the search key value is then retrieved. Such searching techniques require repeated access to records into the storage mechanism to perform key comparisons. In large storage and retrieval systems, such searching, even if augmented by efficient search procedures such as the binary search, often requires an excessive amount of time due to the large number of key comparisons required.

Another well-known and much faster way of storing and retrieving information from computer storage, albeit at the expense of additional storage, is the so-called "hashing" technique, also called scatter-storage or key-transformation method. In such a system, the key is operated on by a hashing function to produce a storage address in the storage space, called the hash table, which is a large one-dimensional array of record locations. This storage address is then accessed directly for the desired record. Hashing techniques are described in the classic text by D. E. Knuth entitled *The Art of Computer Programming*, Volume 3, Sorting and Searching, Addison-Wesley, Reading, Mass., 1973, pp. 506-549.

Hashing functions are designed to translate the universe of keys into addresses uniformly distributed throughout the hash table. Typical hashing functions include truncation, folding, transposition, and modulo arithmetic. A disadvantage of hashing is that more than one key will inevitably translate in the same storage address, causing "collisions" in storage. Some form of collision resolution must therefore be provided. For example, the simple strategy called "linear probing," which consists of searching forward from the initial storage address to the first empty storage location, is often used.

Another method for resolving collisions is called "external chaining." In this technique, each hash table location is a pointer to the head of a linked list of records, all of whose keys translate under the hashing function to that very hash table address. The linked list is itself searched sequentially when retrieving, inserting, or deleting a record. Insertion and deletion are done by adjusting pointers in the linked list. External chaining is discussed in considerable detail in the aforementioned text by D. E. Knuth, in *Data Structures and Program Design*, Second Edition, by R. L. Kruse, Prentice-

Hall, Incorporated, Englewood Cliffs, N.J., 1987, Section 6.5, "Hashing," and Section 6.6, "Analysis of Hashing," pp. 198-215, and in *Data Structures with Abstract Data Types and Pascal*, by D. F. Stubbs and N. W. Webre, Brooks/Cole Publishing Company, Monterey, Calif., 1985, Section 7.4, "Hashed Implementations," pp. 310-336.

Some forms of information are such that individual data items, after a limited period of time, become obsolete, and their presence in the storage system is no longer needed or desired. Scheduling activities, for example, involve data that become obsolete once the scheduled event has occurred. An automatically-expiring data item, once it expires, needlessly occupies computer memory storage that could otherwise be put to use storing an unexpired item. Thus, expired items must eventually be removed to reclaim the storage for subsequent data insertions. In addition, the presence of many expired items results in needlessly long search times since the linked lists associated with external chaining will be longer than they otherwise would be. The goal is to remove these expired items to reclaim the storage and maintain fast access to the data.

The problem, then, is to provide the speed of access of hashing techniques for large, heavily used information storage systems having expiring data and, at the same time, prevent the performance degradation resulting from the accumulation of many expired records. Although a hashing technique for dealing with expiring data is known and disclosed in U.S. Pat. No. 5,121,495, issued Jun. 9, 1992, that technique is confined to linear probing and is entirely inapplicable to external chaining. The procedure shown there traverses, in reverse order, a consecutive sequence of records residing in the hash table array, continually relocating unexpired records to fill gaps left by the removal of expired ones.

Unlike arrays, linked lists leave no gaps when items from it are removed, and furthermore it is not possible to efficiently traverse a singly linked list in reverse order. There are significant advantages to external chaining over linear probing that sometimes make it the method of choice, as discussed in considerable detail in the aforementioned texts, and so hashing techniques for dealing with expiring data that do not use external chaining prove wholly inadequate for certain applications. For example, if the data records are large, considerable memory can be saved using external chaining instead of linear probing. Accordingly, there is a need to develop hashing techniques for external chaining with expiring data. The methods of the above-mentioned patent are limited to arrays and cannot be used with linked lists due to the significant difference in the organization of the computer's memory.

BRIEF SUMMARY OF THE INVENTION

In accordance with the illustrative embodiment of the invention, these and other problems are overcome by using a garbage collection procedure "on-the-fly" while other types of access to the storage space are taking place. In particular, during normal data insertion or retrieval probes into the data store, the expired, obsolete records are identified and removed from the external chain linked list. Specifically, expired or obsolete records in the linked list including the record to be accessed are removed as part of the normal search procedure.

This incremental garbage collection technique has the decided advantage of automatically eliminating unneeded records without requiring that the information storage system be taken off-line for such garbage collection. This is